LETTER

Investigation on Several Model Selection Criteria for Determining the Number of Cluster

Xuelei Hu and Lei Xu

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
E-mail: {xlhu, lxu}@cse.cuhk.edu.hk

(Submitted on May 11, 2004)

Abstract - Determining the number of clusters is a crucial problem in clustering. Conventionally, selection of the number of clusters was effected via cost function based criteria such as Akaike's information criterion (AIC), the consistent Akaike's information criterion (CAIC), the minimum description length (MDL) criterion which formally coincides with the Bayesian inference criterion (BIC). In this paper we study Bayesian Ying-Yang (BYY) harmony learning for model selection via comparing BYY harmony data smoothing criterion (BYY-HDS) with several typical model selection criteria, including AIC, CAIC, and MDL. We empirically investigate model selection on clustering using all these methods on simulated data sets under different sample sizes and real data sets including the well-known iris data set and a gene expression data set. The results of experiments illustrate that BYY-HDS outperforms other methods, especially for small sample size. CAIC and MDL tend to underestimate the number of clusters, while AIC tends to overestimate the number of clusters especially in the case of small sample size.

Keywords - BYY harmony learning, model selection, data smoothing, clustering

1. Introduction

Clustering, as a generic tool for finding groups or clusters in multivariate data, has found wide application in biology, psychology, and economics. In many cases, the number of clusters is not known a priori, model selection techniques are relied upon to determine the number of clusters based on mixture models [4, 13]. Conventionally, model selection is implemented in two phases. In the first phase, we obtain a set of candidate models by some learning principles (usually by maximum likelihood (ML) learning) for a range of models. In the second phase, we select the appropriate model based on some model selection criterion. Popular examples of model selection criteria include Akaike's information criterion (AIC) [1], the consistent Akaike's information criterion (CAIC) [3], and the minimum description length (MDL) criterion [7, 2] which formally coincides with the Bayesian inference criterion (BIC) [8].

Bayesian Ying Yang (BYY) harmony learning [10] was firstly proposed in 1995 and then systematically developed in past years. Not only a number of existing major learning problems and learning methods are revisited as special cases from a unified perspective, but also a harmony learning theory is developed with a new learning mechanism that makes model selection implemented either *automatically* during parameter learning or *subsequently after* parameter learning via a new class of model selection criteria obtained from this mechanism. Moreover, this BYY harmony learning has motivated three types of regularization, namely a data smoothing technique that provides a new solution on the hyper-parameter in a Tikinov-like regularization, a normalization with a new conscience de-learning mechanism that has a nature similar to the rival penalized competitive learning, and a structural regularization by imposing certain structural constraints via designing a specific forward structure in a BYY system. Specifically, the harmony learning on various specific BYY systems with typical structures lead to various specific learning algorithms as well as the detailed forms for implementing regularization and model selection. Readers are referred to [13] for a recent systematical introduction.

Theoretically, the results of implementing BYY harmony learning model selection criteria in two phases and the corresponding automatic model selection techniques are equivalent. To facilitate comparison with conventional model selection techniques that rely on two-phase style model selection, here we focus on one newly proposed BYY harmony model selection criterion in this paper, named BYY harmony data smoothing learning (BYY-HDS) criterion [12, 13]. It is based on the smoothing regularized ML estimators of parameters [12, 13, 14].

We investigate these methods empirically using three groups of simulated data sets with respect to sample size, type of covariance matrix, and data dimension. Moreover, we demonstrate results obtained from two real world data sets. In implementation, we obtain the ML estimates of model parameters by the expectation-maximization (EM) algorithm [5]. We obtain the smoothing regularized ML estimates of model parameters and smoothing parameter by a smoothed EM algorithm [12, 14]. The study has shown that BYY-HDS method being superior to its counterparts, especially when the sample size is small.

The remainder of this paper is organized as follows. In Section 2, we review the background for the model based clustering, and three typical model selection criteria. In Section 3, we introduce BYY harmony data smoothing learning (BYY-HDS) criterion. Experiments are given in Section 4. Finally we draw a conclusion in Section 5.

2. Conventional Approaches on Selection of the Number of Clusters

Gaussian mixture model based clustering assumes that the data are distributed according to a mixture of Gaussian distributions, denoted by

$$p(x) = \sum_{l=1}^{k} \alpha_l G(x|m_l, \Sigma_l) \tag{1}$$

with $\alpha_l \geq 0, l = 1, ..., k$, and $\sum_{l=1}^{k} \alpha_l = 1$, where and throughout this paper, $G(x|m, \Sigma)$ denotes a Gaussian density with mean vector m and covariance matrix Σ. Let $\theta_k = \{m_1, ..., m_k, \Sigma_1, ..., \Sigma_k, \alpha_1, ..., \alpha_k\}$ to be the set of parameters of the mixture with k components. The task of Gaussian mixture model based clustering is to estimate the parameters and the number k based on a finite number of observations $x_1, x_2, ..., x_n$. Given the number of components k, we can estimate the parameters θ_k according to some learning principle. For the ML learning, we estimate the parameters by maximizing the log likelihood function $L(\theta_k)$ denoted by

$$L(\theta_k) = \ln \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} \ln \sum_{l=1}^{k} \alpha_l G(x_i|m_l, \Sigma_l), \tag{2}$$

which can be effectively implemented by the expectation-maximization (EM) algorithm [5].

The problem that remains is how to select the number of components. The two-phase style cluster number selection can be described as follows. In the first phase, we define a range of values of k from k_{min} to k_{max} which is assumed to contain the optimal k. At each specific k, we estimate the parameters θ_k according to some learning principle. In the second phase, with the results $\hat{\theta}_k, k = k_{min}, ..., k_{max}$ obtained in the first phase, we obtain the estimate of the number of clusters $\hat{k}$ from k_{min} to k_{max} according to

$$\hat{k} = \text{arg}\min_{k}\{J(\hat{\theta}_k, k), k = k_{min}, ..., k_{max}\}, \tag{3}$$

where $J(\hat{\theta}_k, k)$ is some model selection criterion.

Next we consider several frequently used model selection criteria: AIC, CAIC, and MDL. These criteria are based on the maximum likelihood (ML) estimators of model parameters. Generally, these three model selection criteria take the form [9]

$$J(\hat{\theta}_k, k) = -2L(\hat{\theta}_k) + A(n)D(k) \tag{4}$$

where $L(\hat{\theta}_k)$ is the log likelihood Eq. 2 based on the ML estimates of mixture parameters, $D(k)$ is the number of independent parameters in k-component mixture, $A(n)$ is a function with respect to the number of observations. According to [4], for arbitrary means and covariances $D(k) = (k-1)+k(d+d(d+1)/2)$ where d is the dimension of x. If spherical covariances are considered we simply have $D(k) = (k-1) + k(d+1)$. Different approaches lead to different choices of $A(n)$. $A(n) = 2$ for Akaike's information criterion (AIC) [1], $A(n) = \ln n + 1$ for Bozdogan's consistent Akaike's information criterion (CAIC) [3], and $A(n) = \ln n$ for Rissanen's minimum description length (MDL) criterion [7] that formally coincides with Schwarz's Bayesian inference criterion (BIC) [8].

These criteria are derived from different theories. One possible interpretation is that the first term is a measure of lack of fit when the maximum likelihood estimators of the mixture parameters are used, the second term is a measure of model complexity that penalizes the first term due to the unreliability of the first term.

3. BYY Harmony Data Smoothing Learning Criterion

The BYY harmony learning is a general statistical learning framework, first proposed in 1995 [10], from which various model selection criteria and automatic model selection methods have been derived [12, 13]. Specifically we consider the one called BYY harmony data smoothing learning model selection criterion (BYY-HDS) for Gaussian mixture model based clustering as follows [13, 14]:

$$J_{BYY-HDS}(\hat{\theta}_k^h, k) = \sum_{l=1}^{k} \hat{\alpha}_l (0.5 \ln |\hat{\Sigma}_l| + 0.5 \hat{h}^2 Tr[\hat{\Sigma}_l^{-1}] - \ln \hat{\alpha}_l), \tag{5}$$

where $\theta_k^h = \{\theta_k, h\}$, with $\hat{\theta}_k^h$ obtained from the data smoothing regularized ML estimates via a smoothed EM algorithm [12, 13], which alternatingly repeats the following steps:

Step 1 Calculate the posterior probability $\hat{P}(l|x_i)$

$$\hat{P}(l|x_i) = \frac{\hat{\alpha}_l G(x_i|\hat{m}_l, \hat{\Sigma}_l)}{\sum_{l=1}^{k} \hat{\alpha}_l G(x_i|\hat{m}_l, \hat{\Sigma}_l)} \tag{6}$$

for $l = 1, ..., k$ and $i = 1, ..., n$.

Step 2 Update parameters by

$$\hat{\alpha}_l = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(l|x_i), \tag{7}$$

$$\hat{m}_l = \frac{1}{n\hat{\alpha}_l} \sum_{i=1}^{n} \hat{P}(l|x_i) x_i, \tag{8}$$

$$\hat{\Sigma}_l = \frac{1}{n\hat{\alpha}_l} \sum_{i=1}^{n} \hat{P}(l|x_i)(x_i - \hat{m}_l)(x_i - \hat{m}_l)^T + \hat{h}^2 I \tag{9}$$

for $l = 1, ..., k$.

Step 3 Update the smoothing parameter h as follows

$$h_{new} = h_{old} + \eta_0 g(h_{old}), \tag{10}$$

where η_0 is a step length constant and

$$g(h_{old}) = \frac{d}{h_{old}} - h_{old} \sum_{l=1}^{k} \hat{\alpha}_l Tr[\hat{\Sigma}_l^{-1}] - \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{i,j} \|x_i - x_j\|^2}{h_{old}^3} \tag{11}$$

with

$$\gamma_{i,j} = \frac{e^{-0.5 \frac{\|x_i - x_j\|^2}{h_{old}^2}}}{\sum_{i=1}^{N} \sum_{j=1}^{N} e^{-0.5 \frac{\|x_i - x_j\|^2}{h_{old}^2}}}. \tag{12}$$

This algorithm not only prevents the covariance matrices from being singular which usually occurs in the EM algorithm on a small size of samples but also provides a new way to update the smoothing parameter. If we let $h = 0$ then this criteria is equivalent to the criterion $J_2^g(k)$ Eq.24 in [11]. Actually, BYY-HDS is an extension of the criterion proposed in [11] for dealing with the small sample size problems.

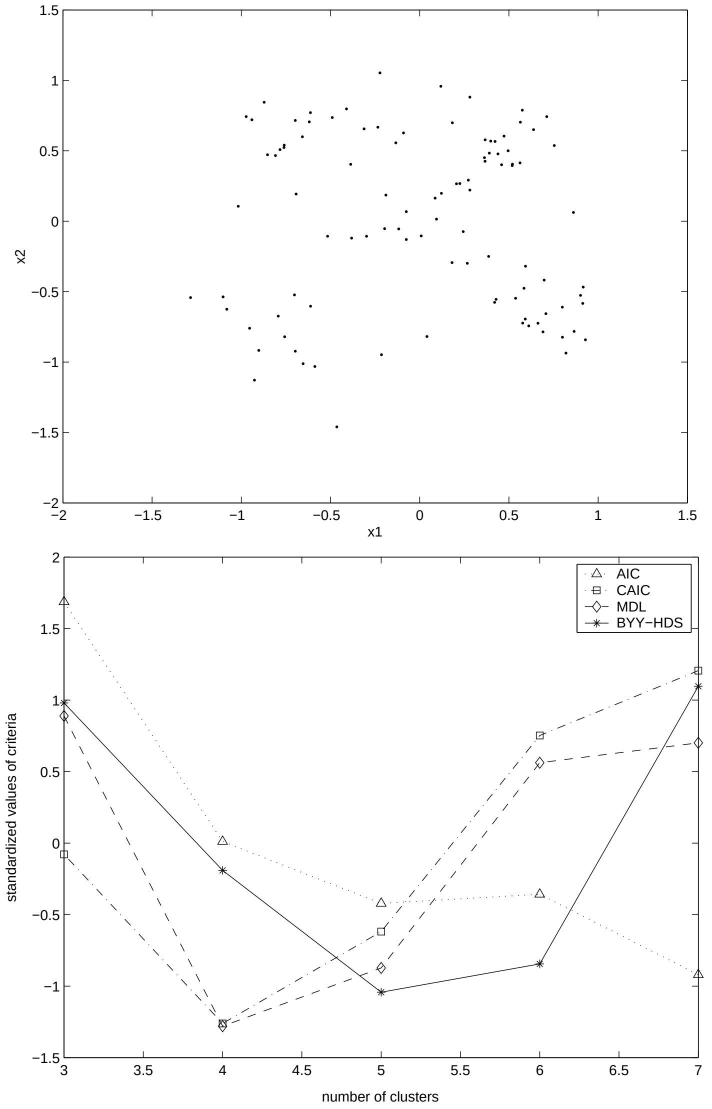

Figure 1. 100 observations generated from 5 elliptic Gaussians (top) and corresponding curves of normalized values of the criteria AIC, CAIC, MDL, and BYY-HDS (bottom)

4. Experimental Comparison

In this section, we investigate the experimental performances of the model selection criteria: AIC, CAIC, BIC, and BYY-HDS on both synthetic data sets and real world data sets. We used the EM algorithm to estimate the mixture parameters for AIC, CAIC, and MDL, and we used the smoothed EM algorithm to estimate parameters for BYY-HDS. The initial parameter estimates for the EM algorithm and smoothed EM algorithm were obtained by randomly allocating observations to sub-populations and computing the prior, sample means and covariance matrices of these initial components. Implemented with the five random starts, the one which gave the largest value of the log-likelihood was used as the solution. The smoothing parameter h for the smoothed EM algorithm was initialized by $h^2 = \frac{1}{dn^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\|x_i - x_j\|^2$.

4.1 Simulated Data Sets

We design three groups simulation experiments to illustrate the performance of each criterion on the data sets with different sample sizes, different types of covariance matrices, and different data dimensions. The observations

Table 1. Rates of underestimating (U), success (S), and overestimating (O) by each criteria on the simulation data sets in 100 replications

	Sample	AIC			CAIC			MDL			BYY-HDS		
Example	size	U	S	O	U	S	O	U	S	O	U	S	O
Spherical	80	0	26	74	69	31	0	48	52	0	11	76	13
	200	0	48	52	16	79	5	12	85	3	6	84	10
	400	0	43	57	12	87	1	8	90	2	5	88	7
Elliptic	100	0	21	79	87	13	0	82	18	0	16	61	23
	250	0	34	66	69	31	0	57	43	0	14	59	27
	500	0	23	77	41	59	0	37	62	1	12	69	19
High Dimensional	100	0	27	73	39	48	13	25	51	24	23	55	22
	500	0	45	55	32	57	11	27	60	13	17	71	12
	1000	0	47	53	10	76	14	8	81	11	8	84	8

are randomly generated from the designed models. Each simulation is repeated 100 times, and model selection procedure is implemented over the 100 replications. The rates of underestimating, success, and overestimating of each methods on simulated data sets are shown in Tab. 1. Due to space limitation, only selected results are shown in figures. To clearly show the curve of each criterion in one figure we normalize the values of each criterion to zero mean and unit variance and then show the normalized values on figures.

4.1.1 Spherical Clustering

In the first example the data sets of size 80, 200, and 400 were randomly generated from a 4-component bivariate Gaussian mixture distribution with equal mixture priors, and equal spherical covariance matrices $0.01I$. We used a Gaussian mixture model with different spherical covariance matrices and specified $k_{min} = 2$ and $k_{max} = 6$.

4.1.2 Elliptic Clustering

In the second example, we considered a more general case of Gaussian mixtures with arbitrary covariance matrices. We randomly generated data sets of size 100, 250, and 500 from a 5-component bivariate Gaussian mixture. We used a Gaussian mixture with arbitrary covariance matrices and set $k_{min} = 3$ and $k_{max} = 7$. The normalized values of each criterion on one simulation with 100 observations and one simulation with 500 observations are shown in Fig. 1 and Fig. 2 respectively. From figures we observe that when the sample size is 100 only BYY-HDS selected the correct number 5, AIC selected the number 7, and CAIC and MDL chose the number 4. When the sample size is 500 all the criteria selected the correct number.

4.1.3 High Dimensional Clustering

In the third example the data sets of size 100, 500, and 1000 were randomly generated from a 4-component 10 dimensional Gaussian mixture distribution with equal mixture priors, and equal spherical covariance matrices $0.036I$. We used a Gaussian mixture model with different spherical covariance matrices and set $k_{min} = 2$ and $k_{max} = 6$.

4.2 Real World Data

In this subsection we investigate the performances of different methods on two real world data sets: iris data set and yeast cell cycle data set. For both the two data sets, the optimum number of clusters is known and the

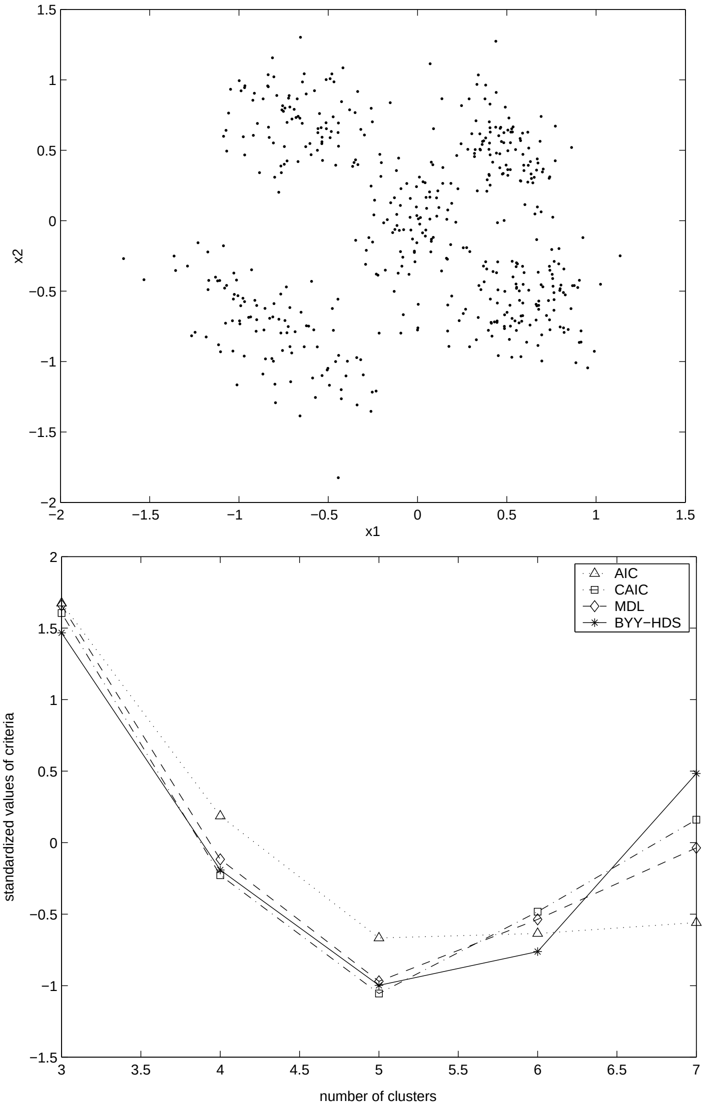

Figure 2. 500 observations generated from 5 elliptic Gaussians (top) and corresponding curves of normalized values of the criteria AIC, CAIC, MDL, and BYY-HDS (bottom)

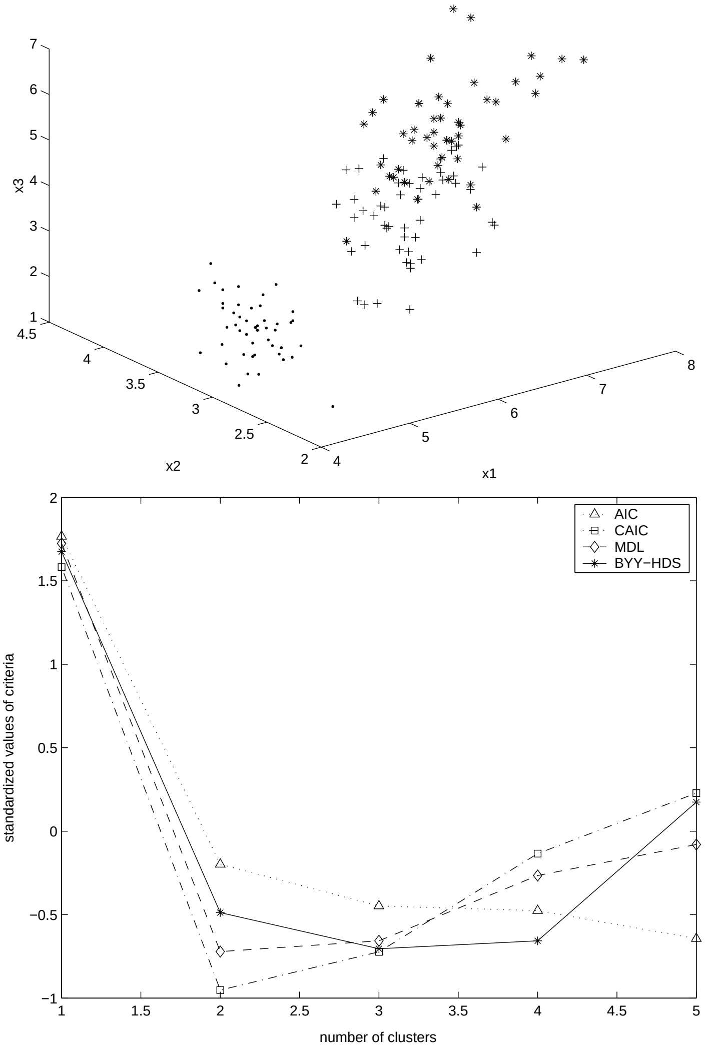

Figure 3. Iris data in first 3-dimensional view (top) and the curves of normalized values of the criteria AIC, CAIC, MDL, and BYY-HDS on iris data set (bottom)

sample size is not large.

4.2.1 Iris Data Set

The well-known iris data set, which was used in [6], contains 150 random samples of flowers from the iris species setosa, versicolor, and virginica ($k = 3$). From each species there are 50 observations for sepal length, sepal width, petal length, and petal width in cm ($d = 4$). Fig. 3 (top) shows the data in first 3-dimensional view. We expect clustering results to approximate this three clusters. We used a Gaussian mixture model with arbitrary covariance matrices because visualization of data shows that the clusters are elliptic in shape. We set $k_{min} = 1$ and $k_{max} = 5$. As shown in Fig. 3 (bottom), AIC chose the number five, CAIC and MDL selected the number two, and BYY-HDS chose the number three which is the correct number.

4.2.2 Gene Expression Data Set

We used the first subset of the yeast cell cycle data in [15]. The data set consists of the expression levels of

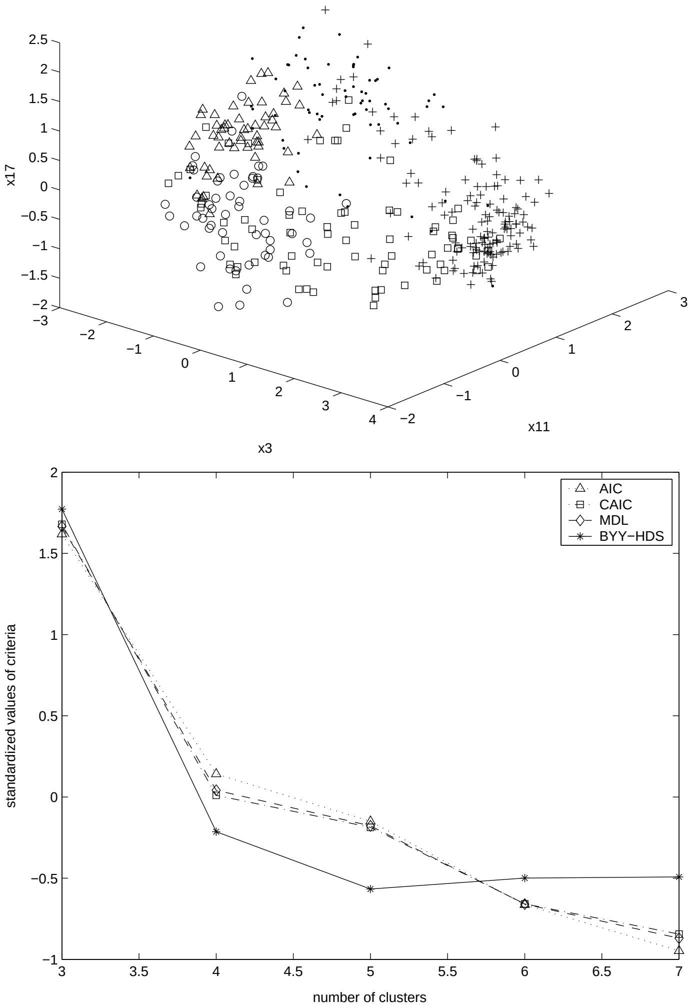

Figure 4. Yeast cell cycle data in 3-dimensional view (top) and the curves of normalized values of the criteria AIC, CAIC, MDL, and BYY-HDS on yeast cell cycle data (bottom)

384 genes ($n = 384$) over 17 time points ($d = 17$). The expression levels of these genes peak at different time points corresponding to the five phases of cell cycle ($k = 5$). We used the normalized data that was shown to be more suitable for clustering in [15]. A 3-dimensional view of data set is shown in Fig. 4 (top). We used a Gaussian mixture model with different spherical covariance matrices because to estimate arbitrary covariance matrices of 17-dimensional data only from 384 observations is difficult and the shapes of clusters shown by visualization are similar spherical. We specified $k_{min} = 3$ and $k_{max} = 7$. As shown in Fig. 4 (bottom), AIC, CAIC, and MDL chose the number 7, and only BYY-HDS selected the correct number 5.

4.3 Discussions

Let us to summarize the main results of the above experiments. Firstly, we measure the performance of the various model selection criteria by their overall success rates. BYY-HDS criterion has the best overall success rate, followed by MDL, CAIC, and AIC. Second, we discuss the properties of these methods with respect to the sample size. BYY-HDS obviously outperforms the other methods for a small sample size. It is reasonable because BYY-HDS uses the data smoothing technique which is a regularization technique that aims to deal with the small sample size problems [14]. While the other methods usually degenerate from their performances in the large-scale sample size. When the sample size increases, these methods get improved accordingly. Finally, we investigate the property of underestimating and overestimating. AIC has high rate of overestimating. CAIC and MDL have a high risk of underestimating the number of clusters especially in the cases of a small sample size. BYY-HDS has no obvious tendency of overestimating or underestimating.

5. Conclusion

We have made an experimental comparison of several cluster number selection criteria based on Gaussian mixture model. The considered criteria include three typical model selection criteria: AIC, CAIC, and MDL/BIC, and BYY-HDS derived from BYY harmony learning. The experimental results show that BYY-HDS is superior to its counterparts, especially when the sample size is small.

Acknowledgment: The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (project No: CUHK 4225/04E).

References

[1] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] A. Barron and J. Rissanen. The minimum description length principle in coding and modeling. *IEEE Trans. Information Theory*, 44:2743–2760, 1998.

[3] H. Bozdogan. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[4] H. Bozdogan. Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan, editor, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 2, pages 69–113, Dordrecht, the Netherlands, 1994. Kluwer Academic Publishers.

[5] A. Dempster, N. Laird, and D.Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Royal Statistical Soc. B*, 39:1–38, 1977.

[6] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[7] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[8] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[9] S. L. Sclove. Some aspects of model-selection criteria. In H. Bozdogan, editor, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 2, pages 37–67, Dordrecht, the Netherlands, 1994. Kluwer Academic Publishers.

[10] L. Xu. A unified learning scheme: Bayesian-Kullback Ying-Yang machine. In D. S. Touretzky, et al, editor, *Advances in Neural Information Processing Systems 8*, pages 444–450. MIT Press, 1996. A part of its preliminary version on *Proc. ICONIP95*, Peking, Oct. 30 - Nov. 3, 1995, 977-988.

[11] L. Xu. Bayesian Ying-Yang machine, clustering and number of clusters. *Pattern Recognition Letters*, 18:1167–1178, 1997.

[12] L. Xu. Bayesian Ying-Yang system and theory as a unified statistical learning approach: (i) unsupervised and semi-unsupervised learning. In S. Amari and N. Kassabov, editors, *Brain-like Computing and Learning*, pages 241–274. Springer-Verlag, 1997.

[13] L. Xu. BYY harmony learning, structural RPCL, and topological self-organizing on mixture models. *Neural Networks*, 15:1125–1151, 2002.

[14] L. Xu. Data smoothing regularization, multi-sets-learning, and problem solving stategies. *Neural Networks*, 16:817–825, 2003.

[15] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

Xuelei Hu is currently a Ph.D. student of the Department of Computer Science and Engineering in the Chinese University of Hong Kong, Hong Kong. She received her M.Phil. degree from the Department of Computer Science, Nanjing University of Science and Technology, Nanjing, China, in 2001. Her research interests include statistical learning and neural computing.

Lei Xu (IEEE Fellow & IAPR Fellow) is a chair professor of Computer Sci. & Eng., Chinese Univ. of Hong Kong. He completed his Ph.D thesis at Tsinghua Univ. by the end of 1986, then joined Dept. Math, Peking University in 1987 first as a postdoc and then was exceptionally promoted to associate professor in 1988. During 1989-93, he worked at several universities in Finland, Canada and USA, including Harvard and MIT. He joined CUHK in 1993 as senior lecturer, became professor in 1996 and took the current chair professor in 2002. Prof. Xu has served or is serving as associate editor for several international journals, including Neural Networks, IEEE Trans. on Neural Networks, as a governor of International Neural Network Society (01-03), the chair of Computational Finance Technical Committee of IEEE Neural Networks Society (01-03), and a past president of Asian-Pacific Neural Networks Assembly. Prof. Xu is known with several well-cited contributions on adaptive PCA and independence learning, classifier combination and mixture model based learning, rival penalized competition, and topological self-organization, as well as Bayesian Ying-Yang unified statistical learning system and theory. Also, he and Oja's invention on Randomized Hough Transform has a wide impact in the field of pattern recognition. He has given over 40 keynote/plenary/invited/tutorial talks in international major neural networks conferences, including ICONIP, WCNN, IEEE-ICNN, IJCNN, etc. He served as a program committee chair of ICONIP'96, ICANNICONIP03, a general chair of IDEAL'98, IDEAL'00, IEEE CIFER'03. He has received several Chinese national prestigious academic awards (including 1993 Chinese National Nature Science Award) and international awards (including 1995 INNS Leadership Award). Prof. Xu is an IEEE Fellow and a Fellow of International Association for Pattern Recognition, and a member of European Academy of Sciences.